View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Xiamen University Institutional Repository

学校编码：10384　　分类号_____密级_____

学号：23020101153065　　UDC_____

厦门大学

硕 士 学 位 论 文

基于选择性集成学习的膜蛋白识别方法研究

The Research of Membrane Protein Prediction based on Ensemble Learning Method

李旭斌

指导教师姓名：邹权 助理教授

专 业 名 称：计算机应用技术

论文提交日期：2013 年　月

论文答辩时间：2013 年　月

学位授予日期：2013 年　月

答辩委员会主席：

评　阅　人：_____ _____

2013 年　月

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：李旭波

2013 年 5 月 29 日

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（ √ ）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：李旭斌

2013年5月29日

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

摘　要

自人类步入后基因组时代，蛋白质组学作为基因组学的下一个重要阶段受到越来越多学者的关注。其中，蛋白质识别和结构预测是蛋白质组学研究的基础环节。目前，生物信息学家开展膜蛋白质识别主要以机器学习分类方法为主，而特征提取和分类算法是其中关键步骤，本文围绕这两点进行了深入地研究。

本文主要研究内容包括：

（1）引入了三种膜蛋白的特征及其提取方法。本文先后引入了指代蛋白质同源信息的 20 维特征；指代氨基酸组成成分-物理化学性质的 188 维特征；指代蛋白质同源信息结合氨基酸在序列中顺序信息的 1000 维特征。实验结果表明，20 维特征具有最高的分类准确率，188 维特征具有最快的提取速度，然而 1000 维特征却没有获得比 20 维更佳的理论结果。

（2）提出了基于最小错分样本交集的选择性集成学习法。本文提出利用最小错分样本交集大小来衡量基分类器间的差异度，从而帮助筛选基分类器。实验结果表明，本文集成分类器在膜蛋白预测上二分类和八分类的准确率分别为 91.2%和 86.1%，和现有最好效果相当，却拥有更高的运行效率。

（3）构建了新的膜蛋白数据集，弥补了已有膜蛋白数据集的不足。发现了参与选择性剪切的多肽中大约 1/3 是膜蛋白。发现了接近 12%的酶具有膜蛋白的特性。开发了基于本文最小错分样本交集的膜蛋白预测平台 BinMemPredict 和选择性集成分类开源工具包 LibSimpleVote。

关键词：膜蛋白；选择性集成学习；最小错分样本交集

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

Abstract

Since life science stepped from genome era into the post genome era, proteomics has attracted growing attention as the next important stage of genomic. Protein identification and classification are the fundamental work of proteomics. Recently, researchers predict membrane protein and its type by machine learning methods. In the whole prediction process, feature extraction and machine learning algorithm are two key components.

The main work of this thesis can be summarized as the following three sections.

(1) Introduce three kinds of features and their extraction methods. We introduce the 20-dimension feature based on position specific scoring matrix(PSSM), the 188-dimension feature based on composition and physicochemical properties of amino acids and the 1000-dimension feature by combining PSSM and order information. The result of experiment show that classification based on the 20-dimension feature achieves the highest accuracy, the 188-dimension feature need the least time in feature extraction, while the 1000-dimension feature do not meet our expectation in accuracy.

(2) Propose selective ensemble learning based on Minimal Intersection of False-Classified-Instance Set(MIFCIS). We introduce MIFCIS concept to evaluate diversity of base classifiers, and combine the base classifiers we found to build ensemble classifier. The result of experiment show that our ensemble classifier achieve accuracy of 91.2% in binary classification and 86.1% in eight-type classification. Our method is more efficient than state-of-the-art methods, and also

厦门大学博硕士论文摘要库

holds high accuracy.

(3) This section can be divided into three parts. Firstly, we build new membrane dataset to remedy deficits of the old dataset. Then, we find about one third of polypeptides evolved in alternative splicing are membrane proteins, and about 12% of enzyme have properties similar to membrane protein. Finally, based on MIFCIS, we develop an online platform(BinMemPredict) to predict membrane protein and release an open-source java library(LibSimpleVote) to handle user-defined prediction task.

Key Words: Membrane Protein; Selective Ensemble Learning; Minimal Intersection of False-Classified-Instance Set

厦门大学博硕士论文摘要库

目录

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

CONTENTS

厦门大学博硕士论文摘要库

厦门大学博硕士论文摘要库

第一章 绪论

1.1 研究背景和意义

1.1.1 蛋白质结构

蛋白质是生命体不可缺少的必要组成部分，也是人类日常生活中必需的营养物质。最早在 1838 年 Gerardus Johannes Mulder 和他的同事们首次发现并描述了蛋白质。之后的科学研究发现，蛋白质几乎参与了细胞生命活动的每个步骤，与各种生命活动息息相关。

蛋白质是由多肽链组成，而多肽链由几十或数百个氨基酸通过脱水缩合形成，因而氨基酸是蛋白质的基本单位。由于氨基酸排列顺序的不同，以及空间立体的不同，形成了蛋白质的多样性结构。如图 1-1 所示，蛋白质结构具体可以分为四级结构。

1） 蛋白质一级结构，指构成蛋白质多肽键的氨基酸序列。

2） 蛋白质二级结构，指在不同氨基酸之间的 C=O 和 N-H 基团间的氢键所形成的稳定结构，主要以α螺旋和β折叠为主。

3） 蛋白质三级结构，指多个蛋白质二级结构元素在三维空间中被折叠为一个蛋白质分子的三维结构。

4） 蛋白质四级结构，指亚基与亚基通过相互作用形成的结构，通常，单独的一个肽链会被称为亚基。

不同的蛋白质结构影响到它的具体生物功能，因此蛋白质的结构多样性造成了不同的生物学功能。想要通过蛋白质一级序列预测蛋白类别，或者想要预测蛋白质的二级结构，首先就要熟悉蛋白质的多级结构。研究蛋白质的结构对了解蛋白质生物功能的实现，以及蛋白质与蛋白质之间的相互作用具有重要意

厦门大学博硕士论文摘要库

义。

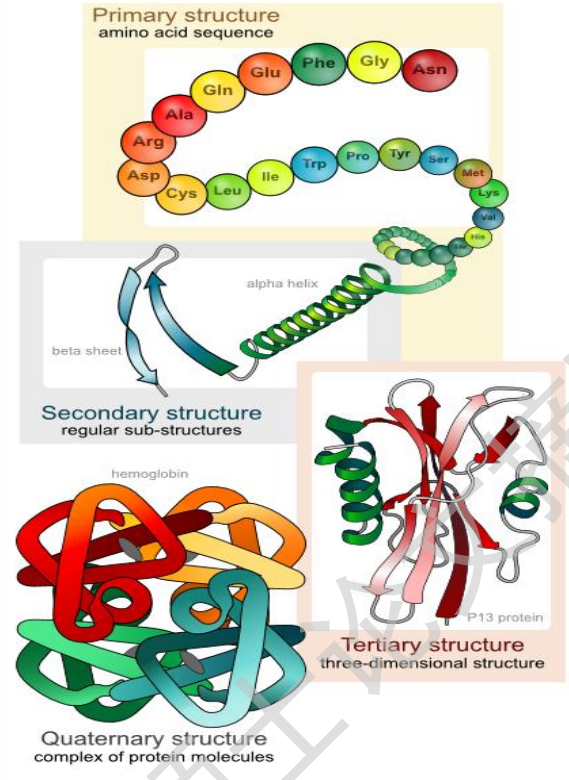

图 1-1 蛋白质一级结构到四级结构的示意图[1]

1.1.2 膜蛋白

生物膜是对细胞内一类膜的统称，包括细胞膜、核膜和线粒体、高尔基体、内质网等细胞器。生物膜系统使得细胞具有一个稳定的内在环境，同时让细胞和周围环境进行物质运输、信息传递、能量交换。膜蛋白是镶嵌在生物膜上的一种结构特殊的球形蛋白质，有的膜蛋白甚至贯穿生物膜，例如跨膜蛋白。作为生物膜功能的执行者，膜蛋白扮演着各种角色，比如受体，通道，膜孔，运载体等等，在细胞中发挥着非常重要的作用。

如图 1-2 所示，膜蛋白又可以被分为两大类八小类。根据膜蛋白和生物膜的结合强度不同，可以将膜蛋白分为两大类：外在膜蛋白(Peripheral Protein)，内在膜蛋白(Integral Protein)。外在膜蛋白只是短时间能和生物膜或者和内在膜

Degree papers are in the “Xiamen University Electronic Theses and Dissertations Database”. Full texts are available in the following ways:
1. If your library is a CALIS member libraries, please log on http://etd.calis.edu.cn/ and submit requests online, or consult the interlibrary loan department in your library.
2. For users of non-CALIS member libraries, please mail to etd@xmu.edu.cn for delivery details.

厦门大学博硕士论文摘要库